B.TECH PROJECT REPORT

Department of Computer Science and Engineering

Indian Institute of Technology Jodhpur

August 2020

PROJECT TITLE	Predicting HashTags from News Networks
TEAM MEMBERS	1. Utpal Gupta(B18CSE058) 2. Yashwant Singh Waskel (B18CSE063)
PROJECT SUPERVISOR	Dr. Suman Kundu

1. ABSTRACT

Social media has shown fast growth in the past few years. Nowadays, social media is not just a place for people to socialise. Instead these days it is also attracting researchers to have a better cognisance about its fundamental working. Real-time micro-blogging services such as twitter use some metadata tags or hashtags(starting with '#') which allows users to apply some tags to their messages or content so that other users can easily find the content in which they are interested. This project mainly focuses on hashtags which can possibly trend on social media in the next few hours. The idea behind the prediction of hashtags works on the fact that online news platforms try to give their readers a quick update on current events and news. And people share their views and opinions on social media and apply some tags to it. Indirectly we're focusing on the essential keywords in the online news articles. So the notion behind the prediction of hashtags is to monitor some esteemed news platforms and extract some keywords from their content and articles that might be relevant to current events and result in future trending hashtags. For that, we apply some state of the art algorithms to those extracted keywords and use some clustering techniques to rank them according to their priorities. After that to find the exact keyword, we can use Twitter's API to extract actual tweets related to those keywords and their hashtags too. Trends on social media are so frequently changed that it's hard to track every one of them and effectively carry out classification of clusters. So we used frequency-based clustering so that results do not vary while new articles are added on news platforms.

2.MOTIVATION

Social media is a place for people to socialise and build relationships, but it also creates a great opportunity for the firms, companies to publicise their products. It can also be used for advertising. So it becomes crucial for these companies to know the current trends on social media to advertise their products correctly. So this requirement of prior knowledge of upcoming trends has motivated us to work on this objective. The objective of this project is to serve as many different ways like sometimes people post aggressive and outrageous content that can lead to violence in society so the prediction of upcoming trends can help police and other people to take early action on such issues.

3.METHODOLOGY

Figure1 summarises the end-to-end framework of this project.

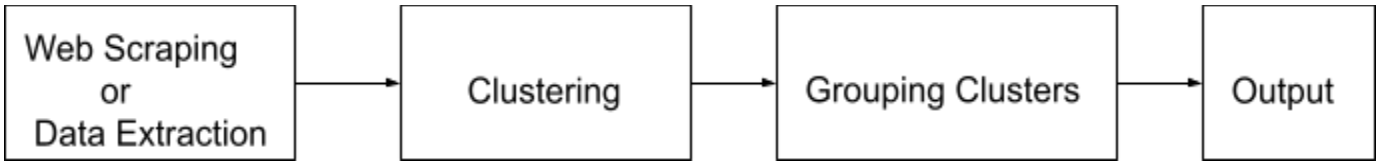

Fig 1. Project design

3.1 Web Scraping or Web Data Extraction

Web scraping or web data extraction is the process of scraping data from a website. Web scraping uses intelligent automation to retrieve hundreds, millions, or even billions of data points from the internet's seemingly endless frontier. It helps to extract all the information on a specific webpage. There are many tools for web scraping that allows parsing the document you have pulled down from the web. One of the tools that we used in this project is BeautifulSoup library in python.

Beautiful Soup is a Python library for getting data out of HTML, XML, and other markup languages. Suppose you've found some webpage whose data you want, but there is no way of downloading it directly. Beautiful Soup helps you pull particular content from a webpage, remove the HTML markup, and save the information. For this process, we first have to request the specific webpage whose data we want to extract, and for that, we use another python library known as requests library.

The requests module allows you to send HTTP requests using Python. The HTTP request returns a response object with all the response data (content, encoding, status, etc.). Now, we can get specific data through HTML tags contents.

With the help of the above, we have successfully extracted meta keywords of each headline from different news websites and collected all the data in the CSV file format. Some of the news websites from where we have extracted the meta keywords are 'Times of India', 'Yahoo News', 'India Today', 'ESPN', etc..
We have collected the news data of the past 4-5 days, including the current day. Now the clustering part comes in action.

3.2 Clustering

Clustering is a set of techniques used to partition data into groups or clusters. Clusters are loosely defined as groups of data objects that are more similar to other purposes in their cluster than they are to data objects in other clusters. There are some essential clustering methods such as hierarchical clustering, density-based clustering and k-means clustering. The clustering method which we used in this project is the K-means clustering method.

3.2.1 K-means Clustering

K-means considers every point in the dataset and uses that information to evolve the clustering over a series of iterations. It works by selecting *k* central points, or *means*. These means are then used as the centroid of their cluster: any point that is closest to a given mean is assigned to that mean's cluster. Once all points are assigned, move through each cluster and take the average of all points it contains. This new 'average' point is the new mean of the cluster. Repeat these two steps until the point assignment stops changing.

K-Means Algorithm

- Specify the number of clusters K.
  1. Randomly initialise K centroids.
  2. Keep iterating until there is no change to the centroids.
  3. Compute the sum of the squared distance between data points and all centroids.
  4. Assign each data point to the closest cluster (centroid).
  5. Compute the centroids for the clusters by taking the average of all data points that belong to each group.

Let's take an example, suppose there is a dataset given as

$$\{2, 3, 4, 10, 11, 12, 20, 25, 30\}$$

Now we want to make two clusters of this dataset i.e., k=2

At first lets randomly take two mean $m_1 = 4$, $m_2 = 12$

Now the two clusters will be

$k_1 = \{2, 3, 4\}$ and $k_2 = \{10, 11, 12, 20, 25, 30\}$

Now, m_1(Mean of k_1) = 3 and m_2(Mean of k_2) = 18

Therefore, $k_1 = \{2, 3, 4, 10\}$ and $k_2 = \{11, 12, 20, 25, 30\}$

Now, $m_1 = 4.75$ and $m_2 = 19.6$

Therefore, $k_1 = \{2, 3, 4, 10, 11, 12\}$ and $k_2 = \{20, 25, 30\}$

Now, $m_1 = 7$ and $m_2 = 25$

Therefore, $k_1 = \{2, 3, 4, 10, 11, 12\}$ and $k_2 = \{20, 25, 30\}$

Now, $m_1 = 7$ and $m_2 = 25$

Thus we are getting the same mean so we have to stop. Therefore, the two clusters for the above dataset are

$$k_1 = \{2, 3, 4, 10, 11, 12\} \text{ and } k_2 = \{20, 25, 30\}$$

3.2.2. Clustering in Python

In python, there is a machine learning library Scikit-learn from where we import TfidfVectorizer. The TfidVectorizer gives the ability to convert words into numbers and we get a table or data frame where each column has a word and each row has a keyword and index are on a left-hand column which has keywords and the words and the count of how many times the word appears within that keyword.
Suppose that we have a group of keywords,

```
keywords = {  'campaign building',
              'ppc campaign generator',
              'how to build ppc campaigns',
              'how do you group keywords',
              'how to group keywords',
              'keyword grouper' }
```

The Dataframe for above keywords which we will get with the help of TfidfVectorizer is given below, and the 'pred' column defines the cluster, i.e., to which cluster the specific keyword belongs to [For clustering we have used K-means clustering method for above keywords and this project].

	build	**campaign**	**generate**	**group**	**grouper**	**keyword**	**ppc**	**pred**
ppc campaign generator	0.00000	0.471964	0.681722	0.000000	0.00000	0.000000	0.559022	0
How to build ppc campaigns	0.764096	0.645102	0.000000	0.000000	0.00000	0.000000	0.000000	0
How do you group keywords	0.000000	0.000000	0.000000	0.764096	0.00000	0.645102	0.000000	1
Campaign Building	0.607144	0.512593	0.000000	0.000000	0.00000	0.000000	0.607144	0

How to group keywords	0.000000	0.000000	0.000000	0.764096	0.00000	0.645102	0.000000	1
Keyword Grouper	0.000000	0.000000	0.000000	0.000000	0.82219	0.569213	0.000000	1

Table - 1

From the above method, we have successfully clustered keywords from different news websites and saved them in the text files according to clusters.

3.3 Output

As we have successfully clustered the keywords now it's time to find the keywords which should be most trending in the present time. The algorithm which we come up with based on data stored is:
A keyword or news which contain it should be the most priority if it is appearing in news media from the past few days including the present day. For example, suppose we have collected the data of last 4 days including present Day(i.e., day1, day2, day3(day before present day), day4(present day)), we have get the keywords whose clusters have been created with their frequency in the respective text files, through this we have got the keywords whose clusters are being created with frequency in a dictionary. So the keywords with most priority will be the data which is occurring on news data collected on day4, day3, day2 and day1 and after this keywords the second priority will be the news keywords which is occurring on day4, day3 and day2 after that third priority will be the news keywords occurring on day4 and day3 and the least priority will be the news occurring only on day4. Now the priority list of dictionaries have been set with keywords and their frequencies. From these we can easily detect the trending keywords and get their hashtags.

Algorithm:

We define some dictionaries to deal with keyword's frequencies.

```
dic1 = {} | dic2 = {} | dic3 = {} | dic4 = {} | dic4321 = {} | dic432 = {} | dic43 = {}
for keyword in dic4:
  if keyword in dic3:
    if keyword in dic2:
      if keyword in dic1:
        if keyword not in dic4321:
          dic4321[keyword]=dic4[keyword] +  dic3[keyword] + dic2[keyword] +dic1[keyword]
```

```
            else:
                dic4321[keyword]+=dic4[keyword] + dic3[keyword] + dic2[keyword] + dic1[keyword]
        else:
            if keyword not in dic432:
                dic432[keyword]=dic4[keyword] +  dic3[keyword] + dic2[keyword]
            else:
                dic432[keyword]=dic4[keyword] +  dic3[keyword] + dic2[keyword]
    else:
        if keyword not in dic43:
                dic43[keyword]=dic4[keyword] +  dic3[keyword]
            else:
                dic43[keyword]=dic4[keyword] +  dic3[keyword]
else:
    continue
```

Through the above algorithm we will get our desired output.

3.4 Exact Hashtag extraction

Exact Hashtags were generated by giving those keywords as a query using Tweepy.

4.RESULTS AND DISCUSSIONS

Here are some important data and results from our project on 6th August 2020.
Time format: HH:MM:SS
From Sports Section:

Time	**Headlines** (from Times of India, India Today etc.)	**Important Keywords**
11:26:50	Rasool: Abbas transformation leaves England outlook cloudy	Pakistan tour of England, England v Pakistan 2020, Mohammad Abbas, England Cricket
11:26:50	IPL Teams Likely Undergo 6 Tests Before They Start	Indian Premier League, UAE

	Training in UAE	
17:13:13	Champions League: Real Madrid coach Zinedine Zidane says Gareth Bale didn't want to play' against Manchester City	Zinedine Zidane, Gareth Bale, Manchester City, Champions League, Real Madrid
17:13:13	IPL 2020: Panic grips BCCI as Vivo likely to exit as Indian Premier League's title sponsor	IPL 2020, BCCI, Vivo, India, Indian Premier League, UAE, Vivo India, China
17:13:13	Fans will see MS Dhoni in his best fighting spirit in IPL 2020: Suresh Raina	MS Dhoni, IPL 2020, UAE Chennai Super Kings
20:26:50	Cristiano Ronaldo Sweats it Out in the Nets Ahead of Juventus vs Lyon, Champions League 2019-20	Cristiano Ronaldo, Champions, League Juventus, Real Madrid
20:26:50	Dobell: Buttler needs to convert after letting Pakistan off the hook	Pakistan tour of England, ICC World Test Championship, England v Pakistan 2020, Jos Buttler, Ben Stokes, England
20:26:50	Jofra Archer bemoans luck with England off-target in the first Test	Pakistan national cricket team, Jofra archer, England vs Pakistan, England cricket team Cricket

Table-2

The above table only represents a small chunk of our data. There were hundreds of headlines and keywords that were extracted from various news platforms. Above given data was still raw. This data was processed further by applying K-means Clustering algorithm and a frequency check. Results are shown in the table below:

Frequency is the number of times a keyword occurred in a cluster.

Keyword	**Frequency**	**Keyword**	**Frequency**
IPL 2020	894	England	766
Cricket	791	Football	444
Pakistan	685	Champions League	353
EngvsPak	503	COVID	324
Badminton	370	league	566
India	326	Dhoni	137

T20WorldCup	261	Ronaldo	324
Real Madrid	78	Juventus	67
Zidane	57	WWE	217

Table-3

Now through these above-mentioned keywords, tweets and corresponding actual hashtags were generated which were on the top of the trending list. Final results are shown below in the table:

Date	Time	Trending Hashtags on Twitter(k=1000)	Related Keywords
06-08-2020	18:00 - 23:00	1. #ENGvsPAK (28.7 k tweets) 2. #JofraArcher(<10k tweets) 3. #IPL2020(<10k tweets)	Pakistan tour of England, England v Pakistan 2020, Jofra archer, IPL 2020, BCCI, Vivo, India, Indian Premier League
07-08-2020	16:00 - 23:00	1. #Ronaldo(163.7k tweets) 2. #ENGvsPAK(10k tweets) 3. #T20WorldCup(<10k tweets) 4. #ChampionsLeague(11.7k tweets) 5. #Lampard(28.8k tweets) 6. #UCLisBack(10k tweets) 7. #Zidane(39.5k tweets) 8. #JuveOL(68.2k tweets) 9. #UnacademyForIPL(<10k tweets) 10. #HalaMadrid(10k tweets)	Zinedine Zidane, Gareth Bale, Manchester City, Champions League, Real Madrid, Cristiano Ronaldo, Real Madrid, Lionel Messi, T20WorldCup, Juventus

Table-4

5.CONCLUSIONS

The keywords were extracted from online news platforms and then ranked according to their priorities and frequency using clustering algorithms. After that, the tweets and corresponding hashtags were extracted using the tweepy library. Results and final Data shows that we were able to predict the upcoming future hashtags and trends.

REFERENCES:

[1] Mateusz Fedoryszak, Brent Frederick, Vijay Rajaram, and Changtao Zhong.2019. Real-time Event Detection on Social Data Streams. In*The 25th ACMSIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19),August 4–8, 2019, Anchorage, AK, USA*.ACM, New York, NY, USA

[2] https://programminghistorian.org/en/lessons/intro-to-beautiful-soup

[3]https://medium.com/predict/three-popular-clustering-methods-and-when-to-use-each-4227c80__ba2b6

[4]https://www.scrapinghub.com/what-is-web-scraping/

NAME AND SIGNATURES OF STUDENTS:

1. Utpal Gupta (B18CSE058)
2. Yashwant Singh Waskel (B18CSE063)

Supervisor: Dr. Suman Kundu